Psychological Review
2011, Vol. 118, No. 1, 155–163

0033-295X/11/$12.00 DOI: 10.1037/a0021774

COMMENT

The Belief Bias Effect Is Aptly Named: A Reply to Klauer and Kellen (2011)

Chad Dube and Caren M. Rotello
University of Massachusetts

Evan Heit
University of California, Merced

In "Assessing the Belief Bias Effect With ROCs: It's a Response Bias Effect," Dube, Rotello, and Heit (2010) examined the form of receiver operating characteristic (ROC) curves for reasoning and the effects of belief bias on measurement indices that differ in whether they imply a curved or linear ROC function. We concluded that the ROC data are in fact curved and that analyses using statistics that assume a linear ROC are likely to produce Type I errors. Importantly, we showed that the interaction between logic and belief that has inspired much of the theoretical work on belief bias is in fact an error stemming from inappropriate reliance on a contrast (hit rate − false alarm rate) that implies linear ROCs. Dube et al. advanced a new model of belief bias, which, in light of their data, is currently the only plausible account of the effect. Klauer and Kellen (2011) disputed these conclusions, largely on the basis of speculation about the data collection method used by Dube et al. to construct the ROCs. New data and model-based analyses are presented that refute the speculations made by Klauer and Kellen. We also show that new modeling results presented by Klauer and Kellen actually support the conclusions advanced by Dube et al. Together, these data show that the methods used by Dube et al. are valid and that the belief bias effect is simply a response bias effect.

Keywords: signal detection, response bias, inductive reasoning, deductive reasoning, belief bias

Supplemental materials: http://dx.doi.org/10.1037/a0021774.supp

Dube, Rotello, and Heit (2010; hereafter referred to as DRH) examined the belief bias effect in syllogistic reasoning using *receiver operating characteristic* (ROC) curves. Visual inspection, area-based tests, and model-based analyses converged to indicate that the interaction between argument validity and conclusion believability that has fueled theoretical debate in the belief bias literature is very likely to be a predictable Type I error. The error was shown to be due to the tacit assumption of a linear ROC relationship in previous analyses of the effect. Because previous theoretical accounts of belief bias (e.g., Dickstein, 1981; Evans, Barston, & Pollard, 1983; Klauer, Musch, & Naumer, 2000; Newstead, Pollard, Evans, & Allen, 1992; Oakhill & Johnson-Laird, 1985; Oakhill, Johnson-Laird, & Garnham, 1989; Polk & Newell, 1995; Quayle & Ball, 2000; Thompson, Striemer, Reikoff, Gunter, & Campbell, 2003) were designed to account for an interaction that is actually a modeling artifact, all of these previous theoretical accounts are now called into question. DRH offered a new interpretation of belief bias, in which subjects are assumed to respond on the basis of a continuously distributed argument strength variable, consistent with the probability heuristics model of Chater and Oaksford (1999) as well as the multidimensional signal detection model of inductive and deductive reasoning advanced by Rotello and Heit (2009; Heit & Rotello, 2010). In the account offered by DRH, believability acts on the positioning of subjects' response criteria: More believable conclusions are associated with a greater willingness to say "valid."

Klauer and Kellen (2011; hereafter referred to as K&K) disputed the response bias account advanced by DRH. Three main arguments are raised: (1) DRH's use of confidence ratings to construct ROCs produced artifactual curvature, because ROCs constructed using binary (yes/no) responses are linear. It follows from this argument that use of the interaction index to assess belief bias may be justified;[1] (2) a new, more appropriate, extension of Klauer et al.'s (2000) *multinomial processing tree* (MPT) model to confidence ratings (henceforth, MPTC) describes the data better than *signal detection theory* (SDT); and (3) the belief bias effect is more than a response bias effect, and in fact DRH's data are consistent with the misinterpreted necessity account (Dickstein, 1981; Newstead et al., 1992).

We show that the evidence K&K presented to support their claims is questionable. In addition, many of the points raised by K&K deal with aspects of MPT modeling that do not clearly relate to the nature of the belief bias effect. When attention is

Chad Dube and Caren M. Rotello, Department of Psychology, University of Massachusetts; Evan Heit, School of Social Sciences, Humanities, and Arts, University of California, Merced.

This research was supported by National Science Foundation Collaborative Research Grant BCS-0616979 to Evan Heit and Caren M. Rotello.

Correspondence concerning this article should be addressed to Chad Dube, Department of Psychology, Box 37710, University of Massachusetts, Amherst, MA 01004-7710. E-mail: cdube@psych.umass.edu

[1] Unless, as Klauer and Kellen (2011, footnote 1, p. 1) pointed out, the ROC has nonunit slope.

shifted from the MPT framework per se, and one instead considers how the available data affect our understanding of belief bias, it can be seen that K&K's analysis actually provides independent support for the account offered by DRH. Moreover, K&K's suggestion that DRH's data are consistent with misinterpreted necessity is unconvincing, and in light of the previous rejection of misinterpreted necessity by Klauer et al. (2000), such discussion does not advance the understanding of belief bias. In what follows, we examine the three basic claims made by K&K. We show how (and to what extent) the data discussed by K&K relate to our understanding of belief bias and conclude that the account offered by DRH remains the only plausible account of the effect.

1. Were DRH's ROC Results an Artifact of the Ratings Procedure? No.

K&K speculated that curvature in ratings-based ROCs of the sort reported by DRH may be a direct consequence of inter- and/or intrasubject variability in response scale usage. In effect, K&K repeated Bröder and Schütz's (2009, p. 587) argument that "the shape of ROCs based on confidence ratings is not diagnostic to refute threshold models [of which Klauer et al.'s (2000) MPT model is one example], whereas ROCs based on experimental bias manipulations are." On this view, the key data are ROCs constructed from independent groups of subjects who make binary yes/no responses under different response bias conditions (e.g., different perceived base rates of valid items). If binary ROCs are linear, but ratings ROCs for the same items are curved, then DRH's conclusions about belief bias are in question. In this section, we show that the evidence for linear binary ROCs is unconvincing.

K&K's support for the assumption of linear binary ROCs rests in part on a recent meta-analysis of recognition memory data (Bröder & Schütz, 2009). There are several reasons to be skeptical of any argument based on the Bröder and Schütz (2009) data. First, pioneering research in signal detection and classification tasks using perceptual stimuli produced curved binary ROCs, which conflicts with the conclusions of Bröder and Schütz (see, e.g., Creelman & Donaldson, 1968; Egan, Schulman, & Greenberg, 1959; Emmerich, 1968; Green & Swets, 1966; Swets, Tanner, & Birdsall, 1961; Tanner, Haller, & Atkinson, 1967). In one particularly elegant study, Egan et al. (1959) compared ratings and binary ROCs within individual subjects who performed a perceptual detection task, finding essentially no difference between them: Both types of ROC were curved for each subject. Though some of these studies were mentioned by Bröder and Schütz (2009), the data were not included in their meta-analysis on the argument that results for perception may not generalize to recognition. This is somewhat at odds with a central assumption made by researchers in perception who applied SDT to recognition in the 1950s and 60s, that is, that similar perceptual mechanisms were involved in both areas (e.g., Egan, 1958; Tanner et al., 1967). Also, more recent findings in reasoning (Dube et al., 2010; Heit & Hayes, in press; Rotello & Heit, 2009) have indicated important similarities across all three domains. For these reasons, the perception data should not be taken lightly.

Although K&K cited the Egan et al. (1959) study, they believed the issue was "moot in recognition," (Klauer & Kellen, 2011, p. 156) because of the study by Bröder and Schütz (2009). Unfortunately, there are limitations to Bröder and Schütz's meta-analysis beyond the selection of individual studies. The analysis involved fitting data from 59 recognition memory experiments using SDT and MPT models. However, in 41 of those 59 data sets, the bias manipulation was varied in only two steps, yielding 2-point ROCs that cannot possibly discriminate linear from curvilinear models.[2] For the remaining 18 data sets, which included more diagnostic 3- and 5-point ROCs, the MPT model was rejected six times according to a compromise chi-square criterion that is appropriate for very large data sets, whereas the SDT model was rejected three times. Moreover, summing the G^2 fit statistics and their corresponding degrees of freedom reveals that the SDT model provided a better overall fit to the data in those 18 data sets ($G^2_{\text{total}} = 332.68$ for SDT vs. 648.98 for MPT). Thus, the data in the meta-analysis indicate that binary ROCs are curved.

There are also serious limitations to the newer results reported by Bröder and Schütz (2009). In all three of their experiments, subjects made binary (yes/no) recognition responses to either word (Experiment 1) or picture stimuli (Experiments 2 and 3). The proportion of studied items on the test list varied over five levels (from 10% to 90%) either between (Experiments 1 & 2) or within subject (Experiment 3). The authors found that, though both models generally provided adequate fits to the data, the SDT model was rejected in one experiment by the standard G^2 criterion, and in most cases a superior fit was obtained with the MPT model. As the authors themselves noted, however, actual base rate manipulations may produce changes in accuracy across conditions (Balakrishnan, 1999; Markowitz & Swets, 1967). All points on an ROC reflect equal accuracy; therefore such effects are at odds with the assumptions of both models. There is also evidence from recognition and perception experiments alike showing changes in the form of the ratings ROC across base rate conditions, which may indicate changes in processing or response strategies (Mueller & Weidemann, 2008; Schulman & Greenberg, 1970; Treisman & Faulkner, 1984; Van Zandt, 2000).[3] Finally, these 5-point ROC results conflict with those reported in the meta-analysis for 5-point ROCs that we discussed previously. For these reasons, firm conclusions cannot be drawn from Bröder and Schütz's newer results.

Although it is clear that an argument based solely on the study by Bröder and Schütz (2009) should be met with skepticism, K&K did report some empirical evidence from reasoning studies to support their claim. In particular, the original MPT model of Klauer et al. (2000; MPTK) and the SDT model were both fit to 10 data sets reported by Klauer et al. Each of these experiments employed a binary response format but varied the perceived base rate of valid syllogisms in three levels across subjects. (Actual base rate did not vary, so it is not obvious whether changes in processing should occur across conditions.) K&K reported that in most cases MPTK outperforms SDT according to the Akaike information criterion (AIC) and the Bayesian information criterion (BIC)

[2] Though Bröder and Schütz (2009) acknowledged this themselves (p. 590), the data in question were nonetheless included in their analysis.

[3] These studies manipulated actual base rates. It is unclear whether the same holds for implied base rate manipulations of the sort employed by Klauer et al. (2000).

fit statistics.[4] In this analysis, some experiments reveal effects of conclusion believability on the reasoning parameters of both models. Although these results seem persuasive, a closer look at the data suggests that a more cautious interpretation is appropriate.

One major issue is that the ROCs in question do not contain enough points to reliably distinguish between MPT and SDT. Although K&K suggested (footnote 6, p. 157) that this renders their results all the more impressive, modeling such restricted data may actually yield misleading or incorrect results. Moreover, AIC and BIC control only for differences in number of free parameters across models, not the key issue of a model's functional form (e.g., Pitt, Myung, & Zhang, 2002).

To assess the difficulty of selecting the better model in this situation, we ran several model-recovery simulation studies.[5] In each study, 1,000 simulated data sets were generated with one of the two models.[6] We selected model parameters completely randomly for the MPT model in one simulation; in the other, the parameters that governed the tendency to guess "valid" were allowed to differ by a maximum of .2. For the SDT model, we also ran two simulations. One assumed that response bias did not differ much across conditions (maximum standardized distance between criteria = .2), as occurred in the data (see Figure 1), and the other allowed more widely spaced criteria (distance up to 1.0). These 4,000 simulated data sets were fit with both models, and the results are shown in Table 1. The distributions of fit differences ($BIC_{MPT} - BIC_{SDT}$) are shown in Figure 2. Overall, these simulations demonstrate that 3-point binary ROCs do not provide good evidence for either model: Both models provided acceptable fits to a given set of simulated data more than 50% of the time, and the distributions of fit differences for the small spacing conditions show most of the data fall near 0. When more widely spaced criteria were used, however, both models claimed their own data at a much higher rate. The model selection challenge may be exaggerated when the operating points are closer together (similar to the Klauer et al., 2000, data) simply because some of those simulated data sets effectively had only one or two unique operating points. In any case, it is clear that even in the widely spaced condition, selection is not easy. These simulations suggest that one should not draw strong model-selection conclusions from fits to 3-point binary ROCs.

A second issue with K&K's reanalysis is that Klauer et al.'s (2000) binary data are unstable, as can be seen in Figure 1. They appear to fall on neither a curve nor a straight line, and there are some large accuracy differences across the implied base rate conditions (e.g., for Experiment 7 the difference is nearly a factor of 2). The stimuli were identical across conditions, so accuracy discrepancies within a given believability condition are troubling. Unfortunately, the MPTK and SDT models do not allow comparison of reasoning parameters across the base rate conditions: Both models would be overparameterized for these data if accuracy parameters for each of the conditions were included. Thus it is unclear whether these changes are significant (note that DRH reported marginal effects of an actual base rate manipulation on accuracy in Experiment 3). In any case, when the accuracy statistic $H - F$ (hit rate minus false alarm rate) is computed separately for each point in these ROCs, the value of the statistic has an average range of .08 for believable problems and .11 for unbelievable. This indicates that changes in accuracy that may be attributed to the believability manipulation could easily be due to random effects of the implied base rate manipulation. Therefore, we do not find it surprising that an interaction between logic and belief appears in some of these experiments and not others (see K&K's Table 5).

Experiment

A better test of the form of binary ROCs in syllogistic reasoning would clearly involve more operating points. Resolving the curvature issue is critical, because K&K's rejection of DRH's conclusions about belief bias depends in large part on their speculation that our ratings ROCs are artifactually curved. The strong curvature in binary ROCs in perceptual tasks (e.g., Egan et al., 1959) and the strong suggestion of curvature in Bröder and Schütz's (2009) meta-analysis (their own conclusions notwithstanding) led us to expect that binary reasoning ROCs would also be curved, contrary to K&K's speculation. To find out, we asked 60 subjects at the University of Massachusetts at Amherst to evaluate the validity of the conclusions of 64 abstract syllogisms. These stimuli were drawn from the same pool used by DRH in Experiment 1 (the structures of which were also used, with believable/unbelievable content, in Experiments 2 and 3). Subjects were randomly assigned to one of five conditions that differed only in the implied base rate of valid problems; all conditions involved 32 valid and 32 invalid problems. This design departs from the manipulation used by Bröder and Schütz but is similar to the one used by Klauer et al. (2000). Specifically, subjects were informed that "at least" or "no more than" 50% of the problems were valid and that they should respond "valid" to approximately 15%, 30%, 50%, 70%, or 85% of the problems, depending on the condition. The phrases "at least" and "no more than" were used for instructed percentages above and below 50%, respectively, and were randomly determined for the 50% group. Subjects were also informed that, though they should produce a certain percentage of "valid" responses, they should still try to be as accurate as possible. To help subjects monitor their performance, we provided feedback on their "valid" response rate (and a reminder of the target response rate) at three different and unpredictable intervals (after 9, 11, or 12 trials). No confidence ratings were collected. In all other respects, the design and procedure were identical to DRH (Experiment 1).

The resulting data are shown as circles in Figure 3. Critically, these binary reasoning ROCs do not appear to be linear, refuting K&K's speculation that the linear nature of reasoning ROCs had been masked by the use of confidence ratings. To be confident of this conclusion, we fit the MPT model proposed by Bröder and Schütz (2009; with parameters relabeled for reasoning rather than

[4] Note that K&K fixed the believability-defined guessing parameter β_b at 1 for MPTK, thus freeing an extra degree of freedom. Although this does not impact the model mathematically (Klauer et al., 2000), we (and DRH) consider it to be psychologically implausible as it results in guessing parameters that reflect ratios of choice probabilities. Moreover, the α parameters, which are supposed to reflect only the base rate manipulation, are reduced in magnitude as a consequence of this restriction. Thus, the resulting tree structure of MPTK is more complex than it appears.

[5] Supplemental materials related to the simulations and new experimental data reported in this study can be found via the supplemental materials link on the first page of the article.

[6] Each operating point was sampled with binomial noise, assuming there were 304 contributing trials (as in DRH, Experiment 2).

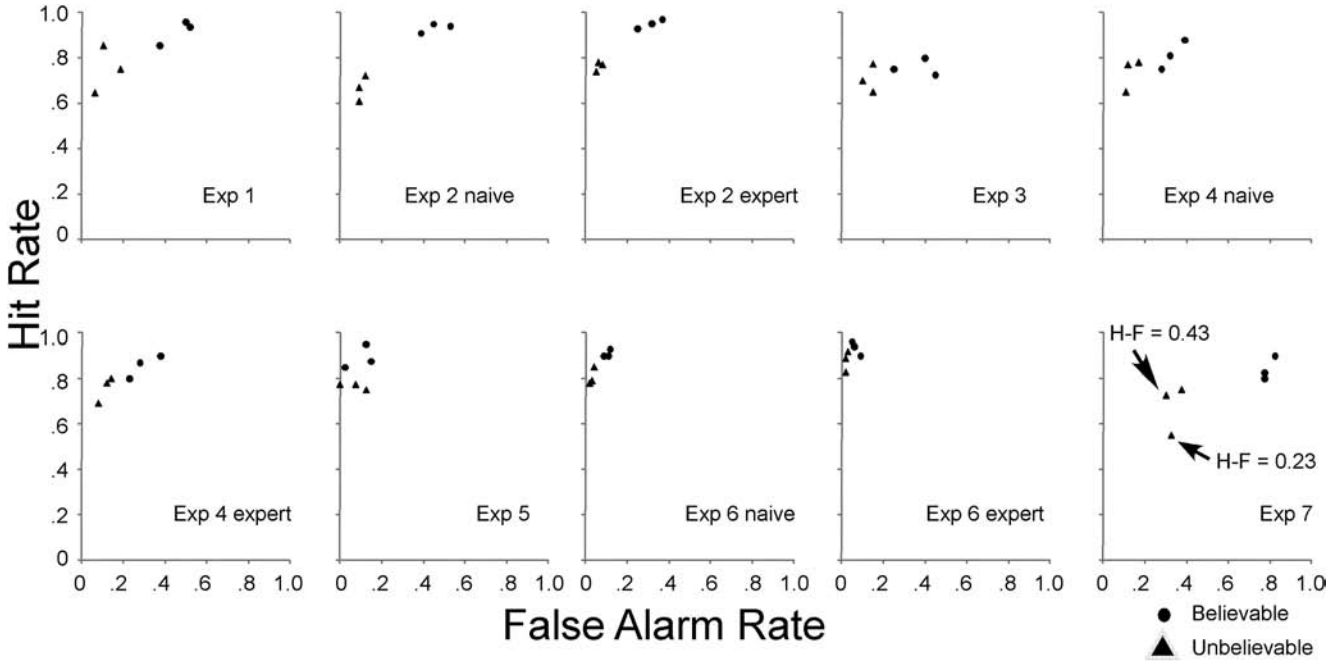

Figure 1. Observed receiver operating characteristics (ROCs) from Klauer et al. (2000), Studies 1–7. H = hit rate; F = false alarm rate.

recognition data), as well as the unequal-variance SDT model used by DRH. Note that, in the present case, no extension of the MPT model (as in the case of ratings data) is necessary. Fit statistics and implied operating points are shown in Figure 3; the best fitting parameter values for both models appear in Table 2. In Figure 3 it can be seen that although both models fit the data well,[7] the SDT model provided a better description of the data than the MPT model. This indicates that binary ROCs for syllogistic reasoning are not linear. Although these models produced a small difference in fit (Burnham & Anderson, 2005), the effect is similar in magnitude to many of the differences reported in K&K's reanalysis of the Klauer et al. (2000) data set (observed AIC difference = 2.26 vs. mean and median AIC difference = 2.02 and 3.07 for K&K's reanalysis). These data replicate DRH's analysis of the three levels of bias in their Experiment 3, as well as the ratings ROCs in all of their experiments, the many binary ROCs reported in the perception literature (e.g., Creelman & Donaldson, 1968; Egan et al., 1959; Emmerich, 1968; Green & Swets, 1966; Swets et al., 1961; Tanner et al., 1967), and the recognition data from the Bröder and Schütz meta-analysis containing more than 2 points. These data indicate that $H - F$ is not an appropriate measure of accuracy in syllogistic reasoning tasks and, therefore, that the interaction index (or, equivalently, a significant analysis of variance interaction) is not an accurate measure of the belief bias effect. Again, we emphasize that because previous theoretical accounts of belief bias were designed to explain findings based on the interaction index (i.e., a threshold model), all of those theoretical accounts are called into question. DRH's conclusions about belief bias were justified and did not reflect distortion due to the use of ratings-based ROCs: Belief bias is a response bias effect, not an accuracy effect.

2. Does the MPTC Analysis Suggest a New Interpretation of Belief Bias? No. MPTC Provides Independent Support for the Model Advanced by DRH

Klauer and Kellen (2011) suggested that our extensions of the MPT model to confidence ratings were inappropriate, as they assumed a particular mapping of internal states to response categories that the authors believed was not justified. They proposed a different extension of the MPT framework, MPTC, which is intended to account for the putative "scale usage artifacts" that the data in Figure 3 suggest are absent. The apparent absence of scale usage effects indicates MPTC may be overparameterized. Nonetheless, we find the MPTC analysis actually further substantiates the analyses and conclusions reported by DRH.

We began by assessing the flexibility of the MPTC and SDT models. On the basis of a simulation in which MPT and SDT models were fit to 3,000 sets of 5 randomly selected points in the

[7] Note that goodness of fit for a given model alone is not relevant in this analysis. Although K&K argued that a poor fit in G^2 indicates a model's parameters should not be interpreted, we (like DRH) are primarily interested in whether the ROC data are curved or linear, because only linear ROCs justify the interaction index as an appropriate measure of the belief bias effect. Hence, the main question is whether a linear (MPT) or nonlinear (SDT) model provides the best fit, not whether a given model provides a perfect description of the data. Our strategy in DRH was to test the parameters of the best fitting model in order to see whether the results agreed with visual inspection and area-based tests of the ROCs. We assume that considering the best model available is preferable to foregoing such an analysis altogether.

Table 1
Model Recovery Simulation Results for 3-Point Binary ROCs

	Generating model			
	MPT		SDT	
No. of cases out of 1,000 simulated data sets	Small bias effect (.2)	Larger bias effect	Small bias effect (.2)	Larger bias effect
MPT fits better	755	720	685	261
MPT fit acceptable[a]	866	923	867	530
SDT fit acceptable[a]	840	762	885	973
Both MPT and SDT fits acceptable[a]	832	747	867	523

Note. ROC = receiver operating characteristic; MPT = multinomial processing tree; SDT = signal detection theory.
[a]Acceptable fits determined by standard chi-square criterion.

upper half of ROC space (Bröder & Schütz, 2009), K&K concluded that SDT models can fit a wider variety of data patterns than MPT models. What that simulation showed, however, was that neither model could fit the random noise that constituted "data": Fit statistics for even the best fit data sets (i.e., those at the 10th percentile) easily exceeded the G^2 criterion for both models. Moreover, that simulation is not relevant for ratings ROCs, for which different models are applied. Analogous to our earlier simulation, we generated simulated belief bias data with the MPTC and SDT models.[8] For the MPT model, we generated 500 data sets assuming $s_h = 1$ (i.e., the rating ROCs were linear) and 500 for which s_h was randomly selected. We also simulated 500 data sets with the SDT model. The results are shown in Table 3, and the distributions of fit differences ($BIC_{MPTC} - BIC_{SDT}$) are shown in Figure 4. As can be seen in Figure 4, the differences in fit clearly favor the MPTC model for MPTC-generated data, with SDT claiming very little of the linear ROC data. When SDT is the generating model, the picture reverses, though MPTC still appears to claim some of the SDT data. This is confirmed in Table 3. When the MPTC model generated curved ROCs, BIC appropriately declared MPTC to have provided the better fit in more than 95% of data sets; less than 3% of the data sets were acceptably well fit ($G^2 < \chi^2$ cutoff) by the SDT model. For the linear MPT-generated ROCs, BIC correctly concluded that MPTC fit better than SDT in 99% of data sets, and, surprisingly, the SDT fit was statistically acceptable in about 20% of cases. These linear ROCs tended to exhibit one of three characteristics: (a) chance level performance, which is linear for both the SDT and MPT models; (b) extremely high accuracy, with all operating points reflecting approximately $H = 1$ or $F = 0$; or (c) near-zero values of either r_v or r_i, which the SDT model captures with its slope parameter.[9] Overall, the SDT model provided a good fit to less than 12% of the MPT-generated ROCs. In contrast, when the SDT model generated the data, 21% of the data sets were well fit by MPTC. The implication of this simulation is that the MPTC model is more flexible than the SDT model.

Given that extra flexibility, the fit advantage claimed by K&K for the MPTC model over the SDT model may simply reflect its greater ability to handle the noise in the data (see Pitt et al., 2002). Even if we take the model at face value, it leads to essentially the same conclusion as DRH: Both models conclude that the effect of believability is on the response bias (or guessing) parameters rather than on accuracy parameters (see Klauer & Kellen, 2011, pp. 159–160). This is as we anticipated in our General Discussion (Dube et al., 2010, p. 854): An MPT model that is capable of producing curved ROCs is likely to reach the same conclusions as the SDT model.

By the same token, MPTC leads to a different conclusion than the original Klauer et al. (2000) model it extends. Specifically, DRH fit the original model (MPTK) to the binary results of Experiment 3, just as K&K did with MPTC for the ratings data in that experiment. The MPTK model revealed an effect of believability on the reasoning parameters, in contrast with MPTC. Thus, one of the two MPT models must be in error, because the "stimulus-state mapping [reflected in the reasoning parameters] is [supposed to be] independent of response format" (Klauer & Kellen, 2011, p. 158).

In sum, the MPTC model developed by K&K (which is meant to control for putative scale usage artifacts) can produce curved ROCs, as does DRH's model, and leads to the same conclusion about the reasoning process as SDT, that the belief bias effect is simply a response bias. This contradicts K&K's own stated conclusions and also contradicts the analysis using Klauer et al.'s (2000) original model, and it reinforces our conclusion that previous theoretical accounts of belief bias, all of which depend on the assumption of linear ROCs, are in question. It also provides independent support for our use of ratings ROCs, because the inclusion of parameters that control for putative scale usage variation did not affect our conclusions regarding the reasoning stage. K&K's MPTC results, together with our new binary ROC data, clearly show that curved syllogistic reasoning ROCs are not an artifact of the data collection method. Most importantly, the MPTC analysis agrees with the conclusions we reached with SDT regarding the belief bias effect: It is purely a response bias effect.

[8] ROCs were generated assuming multinomial noise, with a total of 304 trials of each stimulus type.

[9] A simulation that generated 500 linear ROCs from the MPT model but restricted the reasoning parameters to fall between .2 and .8 yielded only two that were better fit by the SDT model and a dramatically reduced proportion of SDT fits that were statistically acceptable (i.e., 8%).

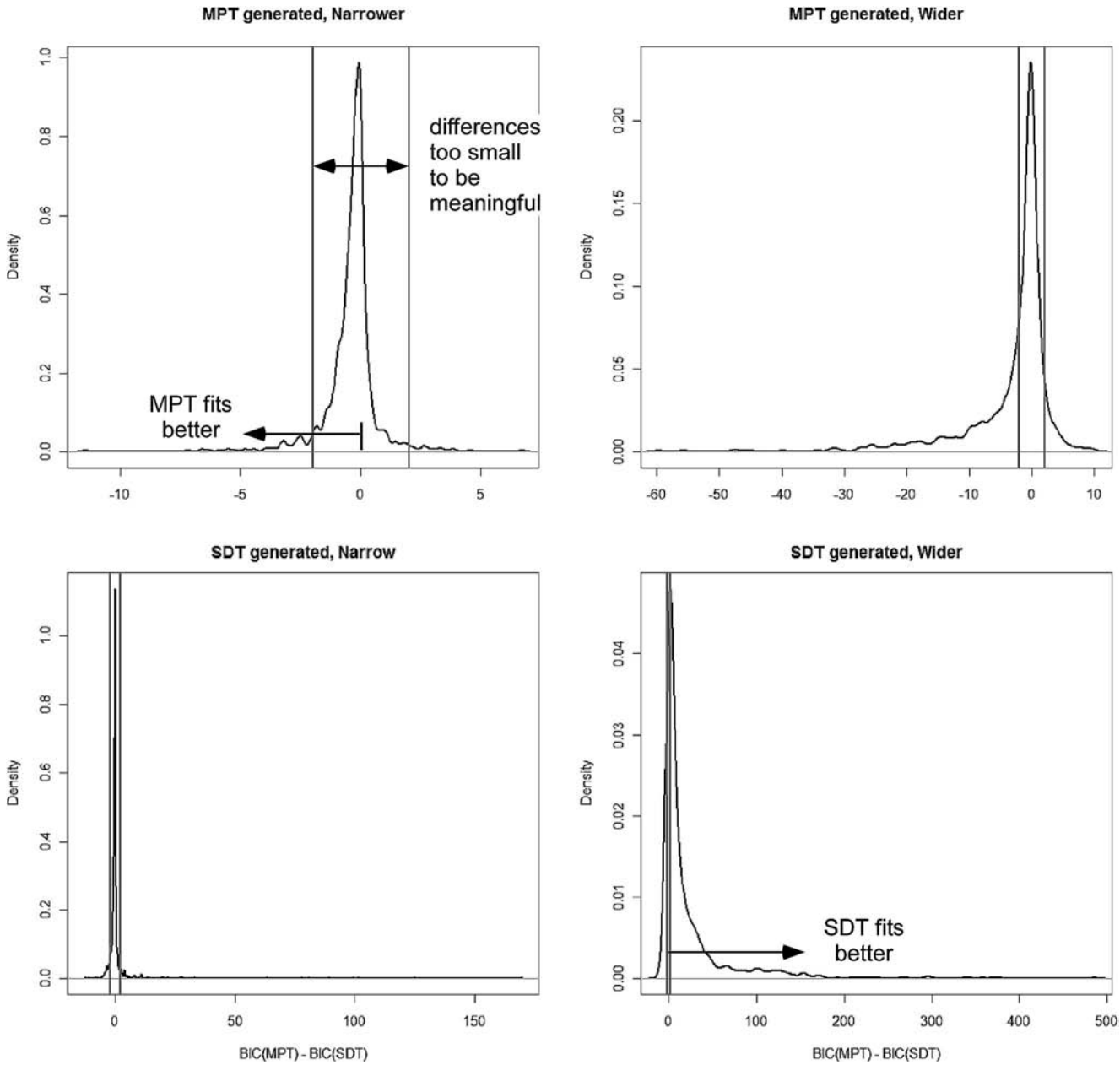

Figure 2. Distributions of change in Bayesian information criterion (ΔBIC) values (multinomial processing tree [MPT] – signal detection theory [SDT]) for fits of the MPT and SDT models to 3-point binary data simulated by MPT (top row) and SDT (bottom row). Small values of ΔBIC that fall within the range –2 to +2 are marked with criteria on *x*.

3. Is Misinterpreted Necessity a Plausible Model of DRH's Data? No.

Klauer and Kellen (2011) suggested that *misinterpreted necessity* (MN; Newstead et al., 1992) provides an adequate account of DRH's data. It is difficult to reconcile this suggestion with the prior rejection of MN by Klauer et al. (2000). After running their own experiments and conducting MPT analyses, Klauer et al. concluded that their pattern of findings "was not consistent with the account by misinterpreted necessity" (p. 870).

K&K's claim with regard to DRH's data is based solely on implications for the reasoning parameters in the MPT model. Both the strict and modified versions of MN differ from other accounts reviewed in DRH in that the effects of belief bias are placed partly in the response stage; thus, MN predicts an interaction in the ROC data that would not necessarily be reflected in the reasoning (*r*) parameters of the MPT model (if

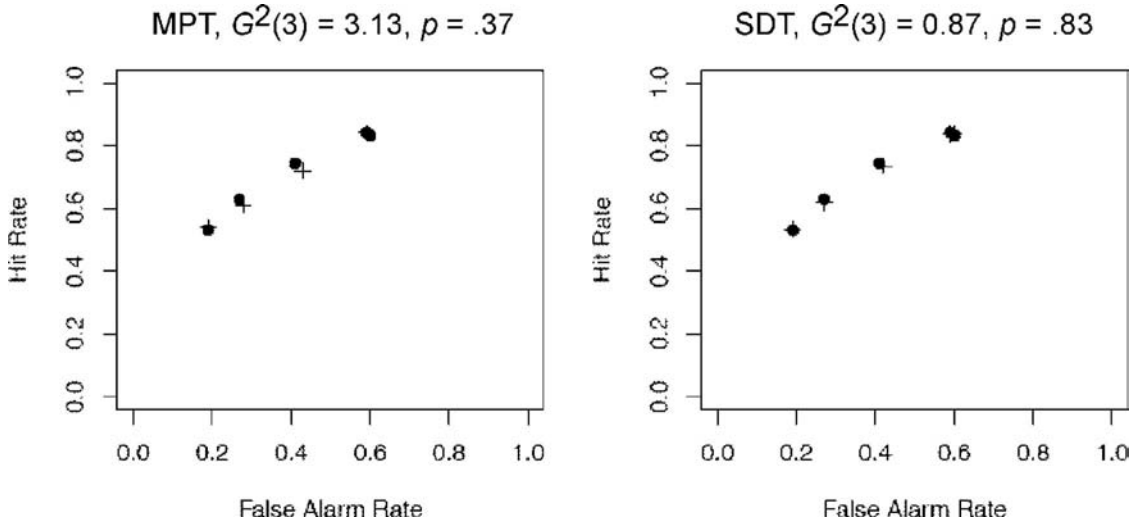

Figure 3. Observed (circles) and model-predicted (crosses) receiver operating characteristics (ROCs) for the multinomial processing tree (MPT; left panel) and signal detection theory (SDT; right panel) models for 5-point binary reasoning data.

such a model were appropriate). Instead, belief bias could affect the bias parameters, β_b and β_u. Repeating arguments from Klauer et al. (2000, p. 858), K&K argued that MN predicts the reasoning parameters conform to the pattern $r_{vb} = r_{vu} = r_v > 0$ and $r_{ib} = r_{iu} = 0$.[10] This assumption requires the MPT model to predict a nonzero interaction index (i.e., a Belief × Logic interaction) unless $\beta_b = \beta_u$.[11] In other words, as noted by K&K, MN predicts the Belief × Logic interaction in $H - F$ without necessitating an effect on the reasoning stage. This by itself is not surprising or informative.

The key question is whether the predictions made by MN hold for reasoning ROCs, which this reply and DRH show are curved. MN predicts reduced false alarm rates for invalid unbelievable problems, yet higher false alarm rates for invalid believable problems (Klauer et al., 2000; Newstead et al., 1992). Because MN focuses on bias effects for invalid problems and predicts relatively little or no influence of believability on the valid problems (for modified and strict versions of MN, respectively), these false alarm rate differences translate to higher accuracy (A_z) for unbelievable than believable problems—in other words, a Belief × Logic interaction. Likewise, Klauer et al. (2000) argued that an absence of a Belief × Logic interaction is evidence against either version of MN.

Our results show that the Belief × Logic interaction that appears when the data are analyzed (i.e., modeled) with the interaction contrast is in fact an error stemming from the contrast itself. This occurs because the contrast assumes (i.e., is a parameter of) an invalid model and is thus not appropriate for the data. When the data are analyzed using an appropriate measure (A_z), the interaction disappears, and the evidence for MN goes away. Although K&K speculated that the analysis of ROC area reported by DRH was unduly affected by the use of ratings data, we have refuted this speculation by showing that binary ROCs for reasoning, like ratings-based ROCs, are curved. Taken together, it is clear that DRH's results are not consistent with MN, because MN predicts an interaction that is an artifact of an unjustified accuracy measure. Moreover, our rejection of MN is consistent with Klauer et al. (2000), who also rejected MN.

Table 2
Best Fitting Parameter Values for MPT and SDT, for 5-Point Binary Reasoning Data

MPT		SDT	
Parameter	Value	Parameter	Value
r_v	.40	μ_v	.98
r_i	.19	σ_v	1.23
α_1	.23	c_1	.88
α_2	.34	c_2	.60
α_3	.53	c_3	.20
α_4	.73	c_4	−.24
α_5	.73	c_5	−.23

Note. MPT = multinomial processing tree; SDT = signal detection theory.

[10] Note that this assumption reduces the MPT to a one-high-threshold model (Macmillan & Creelman, 2005), which implies ROCs differing from those implied by $H - F$ (i.e., it predicts linear ROCs with nonunit slope). Thus the Type I error in $H - F$ could potentially be revealed by this model (in agreement with SDT), even if its assumption of a linear ROC is wrong.

[11] To see that fact, consider the form of the interaction index (DRH, Equation 1): $(H_U - F_U) - (H_B - F_B)$. According to the MPT model, each of the hit and false alarm rates can be written as a simple function of reasoning and bias parameters (see Klauer et al., 2000, Table 2). For example, $H_U = r_{vu} + (1 - r_{vu})\beta_u$. Substituting those functions into the interaction index, and making use of the MN assumptions that $r_{vb} = r_{vu} = r_v > 0$ and $r_{ib} = r_{iu} = 0$, yields a simplified form of the interaction index = $r_v(\beta_b - \beta_u)$. Because $r_v > 0$, the interaction cannot be 0 unless $\beta_b = \beta_u$.

Table 3
Model Recovery Simulation Results for 5-Point Rating ROCs

No. of cases out of 500 simulated data sets	Generating model: MPT (s_h = 1; linear)	Generating model: MPT (s_h random; curved)	Generating model: SDT
AIC: MPT fits better	493	493	11
BIC: MPT fits better	497	476	0
MPT fit acceptable[a]	482	385	105
SDT fit acceptable[a]	100	11	493
Both MPT and SDT fits acceptable[a]	100	8	105

Note. ROC = receiver operating characteristic; AIC = Akaike information criterion; BIC = Bayesian information criterion; MPT = multinomial processing tree; SDT = signal detection theory.
[a]Acceptable fits determined by standard chi-square criterion.

Conclusion

We have shown here and in DRH that the SDT model currently provides the only plausible account of the belief bias effect. It assumes the reasoning process outputs a continuously distributed argument strength value. Subjects are assumed to compare the strength of a given argument to a response criterion. If the argument is strong enough to exceed that criterion, it is accepted; otherwise, it is rejected. Conclusion believability affects the positioning of the response criteria: Believable conclusions result in more liberal responding than do unbelievable ones.

K&K argued that both the SDT and MPT models should be viewed as measurement tools rather than theories (Klauer & Kellen, 2011, pp. 162–163). We believe that this is bad advice. Both SDT and MPT models make strong assumptions about processing, which have historically been the subject of debates concerning, for example, the existence of sensory thresholds (e.g., Swets, 1961) and current debates about the processes that underlie recognition judgments (Parks & Yonelinas, 2009). Though K&K noted that measurement models are not free of assumptions, they nonetheless downplayed the importance of those assumptions because they may not always be useful to discriminate between models, stating that "for such reasons, we prefer to view the simple models used here as measurement tools" (Klauer & Kellen, 2011, p. 162).

Although we appreciate the appeal of simple quantitative tools, applying a particular model to data entails making assumptions, tacitly or otherwise. On this point, we concur with Kinchla (1994, p. 171), who promoted quantitative modeling while nonetheless warning that "an . . . (invalid) mathematical model may lend a spurious precision to the interpretation of the experimental phenomena." We believe that the threshold assumptions of MPT models are violated in perception, memory, and syllogistic reasoning tasks, because the binary and ratings ROCs in these domains are not linear. Consequently, conclusions based on an MPT model may suggest a spurious precision, as they have in the domain of belief bias.

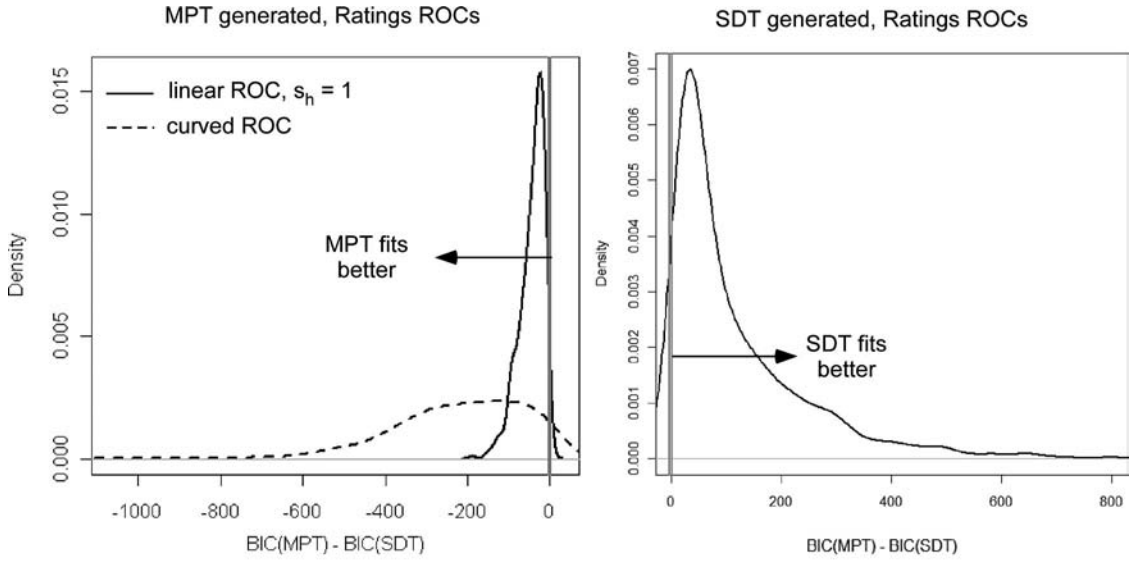

Figure 4. Distributions of change in Bayesian information criterion (ΔBIC) values (multinomial processing tree [MPT] – signal detection theory [SDT]) for the MPTC (an extension of the MPT model to confidence ratings) and SDT models fit to data simulated by MPTC (left panel) and SDT (right panel). ROC = receiver operating characteristic.

To ignore a model's assumptions and view MPT (or SDT) models as mere measurement tools is to commit the same error that has been committed in previous studies of belief bias and recognition memory (see the General Discussion of Dube et al., 2010). As we have shown in DRH, inappropriate use of the interaction contrast in previous analyses of the belief bias effect produced a Type I error that inspired three decades of misdirected theoretical work, explaining a Belief × Logic interaction in syllogistic reasoning that is not really there. We hope, at the very least, that future research will devote attention to the validity of assumptions made in the measurement of the belief bias effect.

References

Balakrishnan, J. D. (1999). Decision processes in discrimination: Fundamental misconceptions of signal detection theory. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 1189–1206. doi:10.1037/0096-1523.25.5.1189

Bröder, A., & Schütz, J. (2009). Recognition ROCs are curvilinear—or are they? On premature arguments against the two-high-threshold model of recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 587–606. doi:10.1037/a0015279

Burnham, K. P., & Anderson, D. R. (2005). *Model selection and multimodel inference* (2nd ed.). New York, NY: Springer.

Chater, N., & Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning. *Cognitive Psychology, 38,* 191–258. doi:10.1006/cogp.1998.0696

Creelman, C. D., & Donaldson, W. (1968). ROC curves for discrimination of linear extent. *Journal of Experimental Psychology, 77,* 514–516. doi:10.1037/h0025930

Dickstein, L. S. (1981). Conversion and possibility in syllogistic reasoning. *Bulletin of the Psychonomic Society, 18,* 229–232.

Dube, C., Rotello, C. M., & Heit, E. (2010). Assessing the belief bias effect with ROCs: It's a response bias effect. *Psychological Review, 117,* 831–863. doi:10.1037/a0019634

Egan, J. P. (1958). *Recognition memory and the operating characteristic* (USAF Operational Applications Laboratory Technical Note No. 58–51). Bloomington, IN: University Hearing and Communication Laboratory.

Egan, J. P., Schulman, A. I., & Greenberg, G. Z. (1959). Operating characteristics determined by binary decisions and by ratings. *The Journal of the Acoustical Society of America, 31,* 768–773. doi:10.1121/1.1907783

Emmerich, D. S. (1968). ROCs obtained with two signal intensities presented in random order, and a comparison between yes–no and rating ROCs. *Perception & Psychophysics, 3,* 35–40.

Evans, J. St. B., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition, 11,* 295–306.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* Oxford, England: Wiley.

Heit, E., & Hayes, B. K. (in press). Predicting reasoning from memory. *Journal of Experimental Psychology: General.* doi:10.1037/a0021488

Heit, E., & Rotello, C. M. (2010). Relations between inductive reasoning and deductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 805–812. doi:10.1037/a0018784

Kinchla, R. A. (1994). Comments on Batchelder and Riefer's multinomial model for source monitoring. *Psychological Review, 101,* 166–171. doi:10.1037/0033-295X.101.1.166

Klauer, K. C., & Kellen, D. (2011). Assessing the belief bias effect with ROCs: Reply to Dube, Rotello, and Heit (2010). *Psychological Review, 118,* 155–164. doi:10.1037/a0020698

Klauer, K. C., Musch, J., & Naumer, B. (2000). On belief bias in syllogistic reasoning. *Psychological Review, 107,* 852–884. doi:10.1037/0033-295X.107.4.852

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.

Markowitz, J., & Swets, J. A. (1967). Factors affecting the slope of empirical ROC curves: Comparison of binary and rating responses. *Perception & Psychophysics, 2,* 91–97.

Mueller, S. T., & Weidemann, C. T. (2008). Decision noise: An explanation for observed violations of signal detection theory. *Psychonomic Bulletin & Review, 15,* 465–494. doi:10.3758/PBR.15.3.465

Newstead, S. E., Pollard, P., Evans, J. S., & Allen, J. (1992). The source of belief bias effects in syllogistic reasoning. *Cognition, 45,* 257–284. doi:10.1016/0010-0277(92)90019-E

Oakhill, J. V., & Johnson-Laird, P. (1985). The effects of belief on the spontaneous production of syllogistic conclusions. *The Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 37*(A), 553–569.

Oakhill, J., Johnson-Laird, P. N., & Garnham, A. (1989). Believability and syllogistic reasoning. *Cognition, 31,* 117–140. doi:10.1016/0010-0277(89)90020-6

Parks, C. M., & Yonelinas, A. P. (2009). Evidence for a memory threshold in second-choice recognition memory responses. *Proceedings of the National Academy of Sciences, USA, 106,* 11515–11519.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109,* 472–491. doi:10.1037/0033-295X.109.3.472

Polk, T. A., & Newell, A. (1995). Deduction as verbal reasoning. *Psychological Review, 102,* 533–566. doi:10.1037/0033-295X.102.3.533

Quayle, J., & Ball, L. (2000). Working memory, metacognitive uncertainty, and belief bias in syllogistic reasoning. *The Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 53*(A), 1202–1223. doi:10.1080/02724980050156362

Rotello, C. M., & Heit, E. (2009). Modeling the effects of argument length and validity on inductive and deductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 1317–1330. doi:10.1037/a0016648

Schulman, A. I., & Greenberg, G. Z. (1970). Operating characteristics and a priori probability of the signal. *Perception & Psychophysics, 8*(A), 317–320.

Swets, J. A. (1961, July 21). Is there a sensory threshold? *Science, 134,* 168–177. doi:10.1126/science.134.3473.168

Swets, J. A., Tanner, W. P., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review, 68,* 301–340. doi:10.1037/h0040547

Tanner, T. A. J., Haller, R. W., & Atkinson, R. C. (1967). Signal recognition as influenced by presentation schedules. *Perception & Psychophysics, 2,* 349–358.

Thompson, V. A., Striemer, C. L., Reikoff, R., Gunter, R. W., & Campbell, J. I. D. (2003). Syllogistic reasoning time: Disconfirmation disconfirmed. *Psychonomic Bulletin & Review, 10,* 184–189.

Treisman, M., & Faulkner, A. (1984). The effect of signal probability on the slope of the receiver operating characteristic given by the rating procedure. *British Journal of Mathematical and Statistical Psychology, 37,* 199–215.

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 582–600. doi:10.1037/0278-7393.26.3.582

Received June 16, 2010
Revision received September 17, 2010
Accepted September 17, 2010 ■